REMARKS

This amendment responds to the Office Action mailed January 6, 2009. In the Office Action the Examiner:

- rejected claims 12-20 under 35 U.S.C. 101 as being directed to non-statutory subject matter;
- rejected claims 12-17, 40, 42-48 and 50-55 under 35 U.S.C. § 103(a) as being unpatentable over Meyerzon et al. (US 6,547,829) in view of Leuski ("Evaluating document clustering for interactive information retrieval");
- rejected claims 18-20, 37-39 and 56-58 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. in view of Leuski and further in view of Rujan et al. (US 6,976,207); and
- rejected claim 49 under 37 U.S.C. §103(a) as being unpatentable over Meyerzon et al. in view of Leuski and further in view of Lambert et al. (US 2002/0038350).

After entry of this amendment, the pending claims are: claims 12-20, 37-40 and 42-58.

REMARKS

Claim Amendments

Applicants have amended Claims 12 and 18 to state that the methods are computer-implemented and that the operations are performed at a server having one or more processors and memory. These amendments are supported by at least paragraphs [0028], [0031], [0036], [0050], and [0051], and Figures 2 and 3.

Applicants have amended Claims 12, 18, 37, 40, 44, 45, 50, and 56 to more fully clarify the usage of "document identifier." Specifically, applicants have changed "document identifier" to "document content identifier," and added language in the independent claims explaining that a document content identifier identifies the content of the documents. That is, two documents with the same document content identifier have the same content, while two documents with different document content identifiers have different content. These amendments are supported by at least paragraphs [0003], [0006], [0007], [0045], [0047], [0048], 0052], [0063], [0066], and [0067], and Figures 5 and 7.

Rejection of Claims 12 – 20 Under 35 U.S.C. § 101

The methods of claims 12 – 20 are performed on a computer, and transform the underlying data structures. Applicants have amended independent claims 12 and 18 to clearly articulate that the methods are performed on a computer. In particular, the operations are performed at a server having one or more processors and memory. The operations identified in the claims are both tied to a particular apparatus and transform the data "to a different state or thing." See *Parker v. Flook*, 437 U.S. 584, 589 n.9 (1978); see also *In re Bilski*, 545 F.3d 943, 956 (Fed. Cir. 2008). The data is transformed by at least: constructing a plurality of tables; storing information identifying documents; updating the information stored; and indexing the representative document. Thus, even applying the more limited "machine or transformation test" articulated recently by the Federal Circuit, Claims 12 – 20 are statutory subject matter under § 101 because they transform the underlying data on a computer.

The rejection of Claims 12 – 20 under 35 U.S.C. § 101 should therefore be withdrawn.

Rejection of Claims 12 – 20, 37 – 40, and 42 – 58 Under 35 U.S.C. § 103

Applicants agree with the Examiner that Meyerzon, Cho, and Wang do not disclose "determining a representative document for the newly crawled document and the identified set of documents … such that at least some of the newly crawled documents are determined to be representative documents and are indexed." Office Action dated 01/06/2009 at page 15.

Further, there are multiple reasons why the newly cited Leuski reference does not teach these missing limitations. First, while Leuski teaches the concept of having a representative document for a cluster of documents, when the representative document of a cluster changes in Leuski, the new representative is not indexed as required in the present claims. Second, the documents in Leuski's document clusters do not all have a same document content identifier, because they don't all have the same content. Third, there are no "newly crawled documents" in Leuski because Leuski teaches search strategies, not crawling new documents on a network. Fourth, there is nothing in Leuski that even corresponds to a "newly crawled document." Leuski just "reorders the documents in the retrieved set" based on "relevance feedback" from the user examining documents. (*See* Leuski, section 4.) Fifth, there is no motivation to combine Leuski with Meyerzon because Meyerzon itself has a method "with the advantage of not having multiple copies of the same document to choose

from." See Office Action dated 01/06/2009 at page 6. Meyerzon achieves that goal with its "first copy wins" methodology. Moreover, Leuski does not address duplicate documents, or a web crawling process, making any combination with Meyerzon questionable.

A. **Background**

Although users interact with a search engine to find relevant information, the search engine must start the process much earlier by crawling the web to collect documents. This earlier phase – crawling the web – is the subject matter of the present application as well as the cited Meyerzon reference. Leuski, on the other hand, addresses the later phase – presenting information to a user based on a particular user query. The crawling phase involves collecting documents, identifying duplicate documents, and indexing certain documents to make them available for later searches. In the crawling phase there is no "user." The search phase involves selecting and ordering documents to present to a user based on a search request by the user.

Fundamentally, Leuski does not teach the limitations in the present application because Leuski addresses the search phase rather than the crawling phase. Leuski's search-phase teachings do not disclose "determining a representative document for the newly crawled document and the identified set of documents … such that at least some of the newly crawled documents are determined to be representative documents and are indexed" as claimed in the present crawling-phase patent application. Applicants explain four reasons for Leuski's failure to disclose below.

B. **When a Leuski representative document changes, the new representative is not indexed as required in the present claims.**

The claims in the present application require

> indexing the representative document when the representative document is the newly crawled document

Claim 12, lines 15 – 16. Further, the claims require repeating the receiving, reading, updating, determining, and indexing operations

> such that at least some of the newly crawled documents are determined to be representative documents and are indexed.

Claim 12, lines 19 – 20. Thus, a representative document does change sometimes, and when it changes the new representative is indexed.

Leuski, in contrast, just reorders the presentation of documents to a user. Leuski, section 4 (page 7). Leuski does not index any documents when the presented documents are reordered. In fact, the terms "index," "indexed," "indexing," etc. do not appear anywhere in

Leuski. Further, it would not make sense for Leuski to index documents because the documents are presumably already indexed. A search engine uses an index to find documents matching a user's query, so unindexed documents would not be returned. Since all of the displayed documents were previously indexed, there would be no reason to index any of them again.

In sum, Leuski does not teach indexing a new representative document, and indexing a new representative document would be inconsistent with Leuski's teachings.

C. Leuski teaches clustering of associated similar documents, not documents with the same document content

Leuski groups together similar documents, not documents with the same content as required by the claims here. The claims first require that:

> documents having the same document content identifier have the same content and documents having different document content identifiers have different content

The claims then specify the operation of forming a set of documents:

> receiving a newly crawled document, such document characterized by a document content identifier and a document rank;
>
> reading information stored in the plurality of tables to identify a set of documents sharing the document content identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

See Claim 12. The grouping in the present claims is thus based on documents having the same content. This is consistent with the title of the application ("Duplicate Document Detection in a Web Crawler System"), and consistent with the stated objectives of the application: "For these reasons, it would be desirable to develop a system and method of detecting duplicate documents crawled by a search engine before the search engine makes any further effort to process these documents" (paragraph [0005]); "Duplicate documents, sharing the same content, are identified by a web crawler system" (paragraph [0006]).

Leuski, in contrast, clusters together documents that are "closely associated." *See* Leuski, section 1, third paragraph. As Leuski explains, "a clustering algorithm brings together similar documents." *See* section 3.3, first paragraph. Moreover, the only example of

clustering in Leuski is provided by Figure 2, in which 50 distinct documents are divided into 14 clusters.

Leuski's clustering of similar documents thus does not teach the grouping of documents with the same content as required by the claims in the present application.

D. There are no "newly crawled documents" in Leuski because Leuski teaches search strategies, not crawling new documents on a network.

The claims in the present application require

> receiving a newly crawled document

Claim 12, line 6. This phrase has a plain and ordinary meaning, which is consistent with the specification. See, e.g., paragraph [0060]. A newly crawled document is a document that a robot / crawler has just downloaded from the network.

The claims in the present application further clarify that the newly crawled document is distinct from documents that are already stored in a search engine database. For example, in Claim 12, after receiving a newly crawled document, the method includes

> reading information stored in the plurality of tables to identify a set of documents sharing the document content identifier of the newly crawled document.

That is, the documents in the database are not newly crawled documents. The specification articulates the same point:

> Upon receiving a newly crawled document, a set of previously crawled documents, if any, sharing the same content as the newly crawled document is identified.

Specification paragraph [0006] (emphasis added).

In contrast, the documents presented to a user in Leuski are all retrieved from the search engine database. The Leuski documents are thus all previously crawled documents. In particular, Leuski identifies the TREC-5 and TREC-6 databases used in the experiments. Leuski, section 5 ("Experimental Setup"). Leuski retrieves documents from the TREC-5 and TREC-6 "Data sets" as shown in Tables 2, 3, 4, and 5.

The searching techniques in Leuski use existing data that was previously collected; at no time does Leuski receive newly crawled documents.

E. There is nothing in Leuski that even corresponds to a "newly crawled document."

Even reading "newly crawled document" more broadly, Leuski does not teach any individual document that is newly added or retrieved. In particular, sections 3.3 and 4 cited by the Examiner refer only to reordering the documents that have already been retrieved from the database:

> A search strategy reorders the documents in the retrieved set at discrete time steps, i.e., when the system receives relevance feedback about examined documents.

Leuski, section 4.

Moreover, the entire goal of the Leuski reference is to organize a retrieved document set so that a user can find the most relevant documents quickly. There is no retrieval of additional documents to modify the list:

> a document organization approach is applied to assist the user in finding the relevant information in the retrieved set

Leuski, section 1, first paragraph.

> In this paper we describe a set of experiments that show the clustering to be a much more effective way of directing a user towards relevant documents among the retrieved set than the ranked list

Leuski, section 1, fifth paragraph.

The Leuski reference thus has no element that corresponds to the newly crawled document in the claims of the present application.

F. There is no motivation to combine Leuski with Meyerzon because Meyerzon itself has a method that identifies and eliminates duplicate documents.

There is no motivation to combine Leuski with Meyerzon because Meyerzon itself has a method "with the advantage of not having multiple copies of the same document to choose from." *See* Office Action dated 01/06/2009 at page 6. Meyerzon implements a "first copy wins" approach as shown in Figure 3 and described at Col. 9, Lines 33 – 40. Further, during a web crawl, as in Meyerzon, there is no "user" to provide "relevance feedback about examined documents," as taught by Leuski.

Fundamentally, Meyerzon addresses elimination of duplicate documents during web crawling, whereas Leuski addresses reordering the presentation of documents retrieved in response to a search query, and performs the reordering based on relevance feedback from a human user. In short, Meyerzon and Leuski are incompatible: Meyerzon addresses finding and eliminating duplicate documents during a web crawl, whereas Leuski teaches reordering documents retrieved in response to a search query; and Meyerzon teaches an automated web

crawler system, whereas Leuski's method requires a human user to provide relevance feedback.

Because Leuski is incompatible with Meyerzon, these references are not properly combined.

CONCLUSION

In light of the arguments presented above, Applicants respectfully request that the Examiner reconsider this application with a view towards allowance. The Examiner is encouraged to call the undersigned attorney at (650) 843-4000 should any issues remain unresolved.

Respectfully submitted,

Date: April 6, 2009 / Gary S. Williams / 31,066

Gary S. Williams
MORGAN, LEWIS & BOCKIUS LLP
2 Palo Alto Square
3000 El Camino Real, Suite 700
Palo Alto, California 94306
(650) 843-4000